DOCUMENT RESUME

ED 295 952 TM 011 695

AUTHOR Jones, J. Patrick; And Others
TITLE A Preliminary Investigation of the Direct Standard Setting Method.
PUB DATE Apr 88
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Certification; *Credentials; Cutting Scores; Health Personnel; *Licensing Examinations (Professions); Minimum Competency Testing; Psychometrics; *Standards; Testing
IDENTIFIERS *Angoff Methods; *Direct Standard Setting Method

ABSTRACT

Three studies assessed the psychometric characteristics of the Direct Standard Setting Method (DSSM). The Angoff technique was also used in each study. The DSSM requires judges to consider an examination 10 items at a time and determine the minimum items in that set a candidate should answer correctly to receive the credential. Nine judges set a passing score for a 100-item portion of a previously administered Examination for the Professional Practice of Psychology (study 1); six judges set passing scores for the 1987 certification examination for operating room nurses in perioperative practice (study 2); and nine judges set passing scores for the 150-item American Board of Health Physics Written Examination (study 3). Data supported the utility of the DSSM as an absolute passing score procedure. The time required for the DSSM was less than that for the Angoff method. In general, DSSM scores were more variable and less reliable than were Angoff cutting scores. The DSSM's reliability may be improved by the use of larger sets; this and other questions merit further investigation. Ten tables are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED295952

A PRELIMINARY INVESTIGATION OF THE DIRECT STANDARD SETTING METHOD

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. PATRICK JONES, GERALD ROSEN
CRAIG SCHOON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

by

J. Patrick Jones, Gerald A. Rosen, and Craig G. Schoon

Professional Examination Service
475 Riverside Drive
New York, NY 10115

TM011695

Paper presented at the annual meeting of the American Educational Research Association in New Orleans, April 1988.

INTRODUCTION

This paper presents the results of three studies undertaken to assess the psychometric characteristics of the Direct Standard Setting Method (DSSM). In each study, judges also used a modified version of the Angoff (1971) technique. After a general description of the DSSM and Angoff method, the procedures used to collect and analyze the passing score data are presented along with the results from each study. The final section of this paper highlights general findings from the three studies and describes future investigations of the DSSM.

DSSM AND ANGOFF PROCEDURES

The DSSM required that judges consider an examination 10 consecutive items at a time, and answer the following question for each set of 10 items, "What is the minimum number of items in this group of 10 that a candidate should answer correctly to receive the credential?" For the DSSM, judges set the minimum test standard in the context of the items in each 10-item set. The item set size of 10 was chosen to allow judges to use proportional estimates (i.e., 7 of 10 equals 70%) and maintain a reasonable memory load while making their ratings.

While orienting judges regarding the use of DSSM, judges were instructed to consider the importance of each item to the credentialing decision, the consequences that lack of knowledge would entail for competent performance, and their estimates of the difficulty level of items. Judges were encouraged to hold candidates more responsible for items that are important to competent performance in the profession and would result in serious harm if the knowledge or skill was not mastered.

DSSM passing scores were determined by computing the sum of a judge's ratings for each of the 10-item sets of the examination. The average of these sums among judges represented the minimum standard for the test.

Angoff passing scores were obtained through the use of a modified Angoff procedure. Judges were first asked to verbalize their notions of the concept of "minimally competent to receive the credential" based on their own opinions and in response to the opinions of other judges. The purpose of this discussion was to foster, to the extent possible, a common notion of minimal competence. Judges were then asked to consider the following question, "What percent of those candidates who are minimally competent for credentialing purposes (i.e., just barely good enough to receive the credential) will answer the item correctly?" To report these percentages, judges indicated on a 11-point scale (i.e., 0%, 10%, 20%, ... 90%, 100%) how many minimally competent candidates would answer the item correctly.

Before the Angoff ratings were collected, judges practiced with a set of items from a previous form of the examination to acquaint themselves with the procedure. The ratings for these practice items were reported to the panel, and high and low raters were invited to express the reasons for their ratings. During these discussions, the difficulty levels (p-values) of the items were announced to provide raters with an idea of how easy or hard the items were for an earlier candidate group.

Angoff passing scores were determined by summing the ratings for each judge. The average of these sums among judges was the minimum passing score.

STUDY I

In this study, 9 judges used the Angoff method and the DSSM to set a passing score for a 100-item portion of a previously administered Examination for the Professional Practice of Psychology (EPPP). The EPPP is a 200-item examination used as part of the licensure process for psychologists in 49 states, the District of Columbia, and 7 provinces in Canada. The 9 judges included 6 members of the Examination Development Committee of the American Association of State Psychology Boards (AASPB) and 3 invited participants.

Following an orientation session, the judges first used the Angoff procedure to set a minimum passing score for the abbreviated form of the EPPP. Once all judges had completed the Angoff ratings, the DSSM was introduced and judges reviewed the same items using the DSSM. Judges recorded both sets of ratings on scannable documents. Passing score data were collected at a workshop convened to examine the feasibility and possible consequences of a suggested national passing score for the EPPP.

To assess the stability of DSSM ratings, judges were asked to use the DSSM to re-rate the same set of test items two months after the conclusion of the workshop. Of the nine judges who attended the workshop, seven judges supplied both sets of DSSM ratings. The second set of DSSM ratings were collected through the mail. During this second rating period, judges also rated the items on a series of item validity scales designed to assess the importance of the item to licensure decisions and the contribution of the item to discrimination between acceptable and unacceptable performance.

Table 1 presents summary statistics for the Angoff and DSSM ratings. The average passing scores obtained using the DSSM and Angoff method were quite comparable. However, the variability of DSSM ratings was greater than Angoff ratings. A positive correlation between item p-values and judges' ratings was evident for both methods, but the correlation was larger for the Angoff method. With respect to item discrimination, the correlation between judges' ratings and the point-biserial correlation coefficients was negative for the Angoff method and weakly positive for the DSSM. This pattern of correlations was also evident for the relationship between judges' passing point ratings and item validity scale and cognitive level ratings.

INSERT TABLE 1 ABOUT HERE

To assess the stability of judges' DSSM ratings, the correlation between workshop ratings and ratings collected 2 months after the workshop was computed. A Pearson product-moment correlation was calculated using the passing scores provided by 7 judges who supplied DSSM ratings on both occasions. This correlation was +.21, and it suggests that DSSM ratings were not consistent across rating sessions.

Inferential statistical analyses performed on the data included an analysis of variance to determine differences among methods, judges, or, for the DSSM, rating sessions, and generalizability analyses (Cronbach, Rajaratnam, & Gleser, 1963) to assess the amount of variance attributable to the various components of each passing score method. The results from a two-way analysis of variance (methods x judges) revealed no significant mean differences among passing score methods (F=2.98,

df=1,9), judges (F=1.26, df=9,72), nor the interaction of methods and judges (F=1.00, df=8,72). An additional two-way analysis of variance (occasions x judges) was conducted using the DSSM data collected at the workshop and through the mail. No significant differences were noted for rating sessions (F=1.06, df=1,9) or judges (F=.88, df=6,54). However, a significant interaction effect was observed (F=2.48, df=6,54; p<.05). For most judges, second session DSSM ratings were slightly lower than their first session ratings. The significant interaction was due to the second session ratings of two judges that were substantially higher than their first session ratings.

Data analysis techniques based on generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) were used to investigate the impact of passing score method, judges, items, and, for the DSSM, passing score session on judges' ratings. Three features of this analytical approach made it especially appropriate for this study. First, generalizability analyses allow one to differentiate among multiple components of rater variance (Brennan & Lockwood, 1980). Second, generalizability analyses can be used to determine the expected variance of judges' ratings when a component of the standard setting model (e.g., judges, items, etc.) is fixed (Colton & Hecht, 1981). These results would be valuable to those individuals who might wish to anticipate in advance of the cutting score session the size of the expected variances associated with different elements of the model. Finally, generalizability analyses yield estimates of the consistency among judges' ratings, and these reliability estimates can be reviewed to compare the extent of rater agreement among passing score methods.

Cronbach et al. (1972) advanced generalizability theory as a means of estimating the variance associated with the facets and objects of measurement of a particular measurement or evaluation model. A facet is the set of all possible conditions of a particular kind. The facets of the present study included judges, items, and, for the DSSM, rating sessions. The ratings supplied by judges for each set of 10 items were the objects of measurement for this study. The sum of the ratings for a 10-item set was chosen as the unit of analysis to permit direct comparisons between the DSSM and Angoff method. The object of measurement was collected according to the facets of the model, and each facet was assumed to have an effect on the object of measurement.

Using a notational system advanced by Brennan (1983), a judge's ratings can be expressed as a linear function of the facets of the model. Equation 1 describes the complete model for this study.

$$Xij = U + Ui\S + Uj\S + Uij\S \qquad (1)$$

The object of measurement, Xij, represents any judge's set of ratings that might be recorded according to the universe of admissible observations included in the model. Equation 1 expresses this set of ratings as a function of the grand mean (U), and the score effects due to items (Ui§), judges (Uj§), and the first-order interaction of these score effects. All of the elements in the model are crossed with each other, and all of the facets are considered to be random effects.

Tables 2 and 3 present the results of the generalizability analyses for the model in Equation 1 for the Angoff method and DSSM, respectively. The largest variance components were due to

the items facet and the interaction of items and judges. The variance components associated with the items facet were larger than the variance components associated with the judges facet. This finding supports the validity of the passing score methods as one would expect that the test items should account for the largest amount of variance among judges' ratings. While the estimated variance components for the judges facet were comparable among methods, the variance component for the interaction effect was substantially larger for the DSSM.

The data presented in the lower portions of Tables 2 and 3 correspond to reliability indices and the estimated variance components for elements of the model with certain facets fixed or controlled for. The reliability indexes for Angoff ratings were considerably larger than reliability indexes for the DSSM. Estimated variance components controlling for certain aspects of the model were larger for the DSSM as compared to the Angoff technique. Thus, these findings suggest that the Angoff method yielded ratings that were more stable and contained less error variance than the DSSM.

INSERT TABLES 2 & 3 ABOUT HERE

Table 4 presents generalizability study results for the random effects model of Equation 1 using the DSSM ratings obtained from 7 judges after the conclusion of the passing score workshop. While the estimated variance components for items and the interaction of items and judges for the DSSM remained larger than the Angoff ratings, the amount of residual error was substantially smaller than the DSSM results obtained at the workshop. The reliability indexes and estimated variances for fixed elements of

the DSSM ratings collected after the workshop were generally better than the results of the first set of DSSM generalizability statistics.

INSERT TABLE 4 ABOUT HERE

Data from the second set of DSSM ratings support an interpretation that the method could demonstrate acceptable psychometric properties. Further investigations of the DSSM were undertaken to isolate the conditions under which the DSSM could be used with confidence regarding its validity and reliability.

STUDY II

This study extended the investigation of the DSSM by collecting judges' ratings of the ability of the method to accurately separate examinees into a group that should receive the credential and a group that should not. In addition, the amount of time required to complete Angoff and DSSM ratings were compared. The context of the credentialing decision was also changed for this study from licensure to certification.

For this study, 6 judges used the Angoff method and DSSM to set passing scores for the 1987 certification examination for operating room nurses in perioperative practice. Passing this 250-item test is a requirement for certification by the National Certification Board: Perioperative Nursing, Inc., and the examination addresses the roles and responsibilities of perioperative nurses in the preoperative, intraoperative, and postoperative phases of patient care and management. The judges were members of the NCB Item Review Committee.

The rating session began with an orientation regarding the use of the modified Angoff procedure and a practice session with a set of retired test items. Judges then rated the examination using the Angoff technique. After a lunch break, judges received instructions on the use of the DSSM and became acquainted with the procedure by using it to rate a set of retired items. The judges then used the DSSM to set a passing score for the test.

Judges recorded passing score ratings on scannable documents. They were also asked to record the time that they began and ended the rating session for each standard setting method. At the conclusion of each technique, judges indicated the amount of confidence that they had that the procedure would accurately separate those candidates who should be certified from those who should not be certified. Rating times and confidence estimates were recorded by judges on a separate biographical data sheet.

Summary statistics for the NCB standard setting activities are presented in Table 5. The mean DSSM passing score was higher than the Angoff standard, and further analyses revealed that this difference was statistically significant (F=71.9, df=1,24; p<.001). With respect to the variability of judges' ratings, the standard deviation and range of average ratings were larger for the DSSM. The amount of time needed to supply DSSM ratings was 50% less than the time required to use the Angoff procedure. The degree of confidence that judges expressed regarding the ability of the methods to accurately separate examinees into certify/don't certify groups was comparable among methods. Weak positive correlations were noted between judges' ratings and item validity scale data for both methods, however, the correlation between

passing score ratings and critical consequences validity scale ratings was substantially larger for the DSSM.

INSERT TABLE 5 ABOUT HERE

In addition to the statistically significant differences in average ratings among methods, the results from a two-way analysis of variance (methods x judges) revealed mean differences for judges (F=5.8, df=5,120; p<.001) and the interaction of methods and judges (F=24.9, df=5,120; p<.001). The significant judges main effect and interaction was due to relatively low DSSM ratings by 2 of the 6 judges.

Generalizability analyses were conducted on the NCB data set using the model presented in Equation 1. Tables 6 and 7 present estimated variance components and reliability indexes for the Angoff method and DSSM, respectively. The estimated variance components for all facets of the DSSM were larger than the estimated variance components for the Angoff method. The reliability of average ratings and the reliability of ratings for a single judge were lower in this study as compared to the EPPP data set. However, the reliability of Angoff ratings was slightly higher than DSSM reliability estimates.

INSERT TABLES 6 & 7 ABOUT HERE

In summary, DSSM passing scores were slightly higher and more variable than Angoff cutting scores. The rating time required by judges to complete the DSSM was nearly 50% less than the time needed to use the Angoff procedure. Judges expressed confidence that the passing scores derived from both methods would accurately separate examinees into certify/don't certify groups. The

reliability of judges' ratings was generally low for both approaches.

STUDY III

This study was conducted to determine if the presentation order of passing score methods may have effected the results of earlier studies. For the first two investigations, DSSM data were collected after lengthy Angoff rating sessions, and judges' DSSM ratings may have been influenced by rater fatigue and other situational factors. Another goal of this study was to examine the psychometric characteristics of the DSSM for a credentialing program outside of the health-care field.

Judges in this study were 9 members of the American Board of Health Physics (ABHP) Test Development Committee. These judges used the DSSM and Angoff method to set passing scores for the 150-item ABHP Written Examination. The ABHP certification program is designed for individuals with a degree in the biosciences and work experience in the field of health physics. Standard setting data were collected as part of a passing point workshop sponsored by the ABHP.

The ABHP workshop began with an introduction to the DSSM and a practice rating session. After reviewing the results from the practice ratings, judges used the DSSM to set a passing score for the ABHP test. After a lunch break judges were instructed on the use of the Angoff procedure and a practice rating session was conducted. The practice ratings were discussed and the judges then rated the ABHP test using the Angoff technique.

Passing score ratings were recorded by judges on scannable forms. The time needed to complete ratings and confidence

estimates were compiled by judges on a separate biographical data sheet.

Descriptive statistics for the ABHP passing point data are presented in Table 8. As was the case in the previous study, the DSSM passing score was higher than the Angoff standard (F=18.8, df=1,8; p<.005). The variability and range of judges' passing scores were slightly larger for the Angoff procedure. The average time needed to rate items using the DSSM was approximately 25% less than the time required to complete Angoff ratings. With respect to the confidence of judges regarding the ability of the procedures to yield accurate certify/don't certify decisions, the average confidence rating was higher for the DSSM. Finally, the correlations between passing score ratings and item difficulty and discrimination indexes were weakly positive for both methods.

INSERT TABLE 8 ABOUT HERE

A method x judges analysis of variance was conducted using the ABHP rating data set. The main effect for methods previously cited was the only statistically significant finding from this analysis of variance.

Generalizability analyses were completed for the Angoff method and DSSM based on the model presented in Equation 1. Tables 9 and 10 present the results from these analyses for the Angoff procedure and DSSM, respectively. The estimated variance component for items was negative for each passing score method. A variance component must be positive by definition. In some instances, however, estimates of variance components are negative (c.f., Brennan, 1983; Brennan & Lockwood, 1980). Cronbach et al. (1972) suggested replacing negative estimated variance components

with zero wherever they occur, while others advise leaving them unchanged (c.f., Sirotnek, 1970). Due to the exploratory nature of this study and the small number of item sets upon which these estimates were based, the estimated variance components for items were left as negative numbers. Reliability estimates were not computed for either method to avoid the logical problem associated with a negative reliability coefficient.

INSERT TABLES 9 & 10 ABOUT HERE

The estimated variance components for items and judges were smaller for the DSSM as compared to the Angoff method. The estimated variance component due to the interaction of items and judges plus residual error was substantially larger for the DSSM. The variance components for average ratings and elements of the model given that certain facets were fixed were comparable among the standard setting approaches.

SUMMARY AND CONCLUSIONS

The findings from the three studies support the utility of the DSSM as an absolute passing score procedure. Judges expressed confidence that the DSSM would accurately categorize examinees into "credential" and "don't credential" groups. The time required to conduct the DSSM rating sessions was substantially less than the time needed for the Angoff method. While the relationship between DSSM ratings and item validity scale data was not as strong as originally anticipated, these correlation coefficients were comparable with, or superior to, those obtained for the Angoff data sets.

In addition to the findings summarized previously, the

biggest differences between the passing score results obtained from the use of the two approaches were the size of the cutting scores, the variability of judges' ratings, and the reliability of ratings. In two of the three studies, average DSSM passing scores were more than 9 points higher than average Angoff standards. This discrepancy may be attributable to the focus of the two standard setting questions. For the DSSM used in these studies, judges indicated the minimum number of items that a candidate should answer correctly to receive the credential. Judges may have responded to this task by prescribing a standard that is higher than the minimum level of proficiency needed to pass the examination. For subsequent investigations of the DSSM, the following question will be posed to judges, "How many of these 10 questions should a candidate answer correctly to be considered 'just good enough' to receive the credential conferred by this test?" The intent of the revised DSSM question is to bring judges' item ratings more in line with current performance characteristics of individuals within a profession.

In general, DSSM passing scores were more variable and less reliable than Angoff cutting scores. Subsequent investigations of the DSSM will address these problems by encouraging judges to review item validity scale data during the rating process and providing judges with general information regarding the difficulty of test items (e.g., easy, medium, hard). This information should serve as a common referent or anchor point for judges and, thus, improve the stability of DSSM ratings.

Other areas of research involving the DSSM include an examination of the psychometric properties of the individual item version of the technique and variants of the item set size used

for the group version. Based on preliminary results with 10-item sets, the use of larger sets may improve the reliability of judges' ratings by increasing the range of set ratings

REFERENCES

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education.

Brennan, R.L. (1983). Elements of generalizability theory. Iowa City, IA: The American College Testing Program.

Brennan, R.L., & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Colton, D.A., & Hecht, J.T. (1981, April). A preliminary report on a study of three techniques for setting minimum passing scores. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scales and profiles. New York: Wiley.

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.

Sirotnek, K. (1970). An analysis of variance framework for matrix sampling. Educational and Psychological Measurement, 30, 891-908.

TABLE 1. Summary Statistics for EPPP Standard Setting Data

	Method	
Statistic	**Angoff**	**DSSM**
Average passing score	68.3	70.6
Standard deviation of passing scores	2.1	4.7
Range of average ratings	63.3 - 71.0	66.0 - 77.0
Correlation of ratings[1] with item p-values	.46	.33
Correlation of ratings with item r_{pbis}	-.20	.14
Correlation of ratings with **importance** validity scale ratings (1 = not important, 2 = medium importance, 3 = very important, 4 = critically important)	-.19	.10
Correlation of ratings with **ability of item to discriminate** validity scale ratings (1 = will not contribute, 2 = small contribution , 3 = substantial contribution, 4 = critical contribution)	-.07	.12
Correlation of ratings with **cognitive level classification** of items (1 = recall/recognition, 2 = interpretation/application/analysis, 3 = synthesis/evaluation)	-.07	.07

[1] Correlations are based on ratings and/or item statistics summed for 10-item sets

Table 2. EPPP Generalizability Analyses for Angoff Method

Source	df	ms	Estimated Variance Component
Items (i)	9	.38544	.03090
Judges (j)	8	.18703	.00797
ij	72	.10731	.10731

r_{kk} = Reliability of average ratings for all judges = 0.72

$\bar{r}_{11}$ = Reliability of ratings for a single judge = 0.22

$\hat{\sigma}^2\ \bar{x}$ = Estimated variance of average ratings = $\hat{\sigma}^2$ (i)/n_i + $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$(ij)/$n_i n_j$ = 0.005

$\hat{\sigma}\ \bar{x}$ = 0.070

$\hat{\sigma}^2\ \bar{x}_r$ = Estimated variance of ratings over sample of raters = $\hat{\sigma}^2$(j) +$\hat{\sigma}^2$ (ij)/n_i = 0.019

$\hat{\sigma}\ \bar{x}_r$ = 0.137

$\hat{\sigma}^2\ \bar{x} \mid j$ = Estimated variance of mean ratings generalizing over samples of 10 sets of items for a fixed set of 9 judges = $\hat{\sigma}^2$(i)/n_i + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .004

$\hat{\sigma}\ \bar{x} \mid j$ = 0.065

$\hat{\sigma}^2\ \bar{x} \mid i$ = Estimated variance of mean ratings generalizing over samples of 9 raters for a fixed set of items = $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .003

$\hat{\sigma}\ \bar{x} \mid i$ = 0.053

Table 3. EPPP Generalizability Analyses for DSSM

Source	df	ms	Estimated Variance Component
Items (i)	9	1.9568	.10309
Judges(j)	8	1.1278	.00988
ij	72	1.0290	1.02901

r_{kk} = Reliability of average ratings for all judges = 0.47

$\bar{r}_{11}$ = Reliability of ratings for a single judge = 0.09

$\hat{\sigma}^2\ \bar{x}$ = Estimated variance of average ratings = $\hat{\sigma}^2$ (i)/n_i + $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$(ij)/$n_i n_j$ = .023

$\hat{\sigma}\ \bar{x}$ = .151

$\hat{\sigma}^2\ \bar{x}_r$ = Estimated variance of ratings over sample of raters = $\hat{\sigma}^2$(j) +$\hat{\sigma}^2$ (ij)/n_i = .113

$\hat{\sigma}\ \bar{x}_r$ = .336

$\hat{\sigma}^2\ \bar{x} \mid j$ = Estimated variance of mean ratings generalizing over samples of 10 sets of items for a fixed set of 9 judges = $\hat{\sigma}^2$(i)/n_i + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .022

$\hat{\sigma}\ \bar{x} \mid j$ = .147

$\hat{\sigma}^2\ \bar{x} \mid i$ = Estimated variance of mean ratings generalizing over samples of 9 raters for a fixed set of items = $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .013

$\hat{\sigma}\ \bar{x} \mid i$ = .112

Table 4. EPPP Generalizability Analyses for Second Set of DSSM Ratings

Source	df	ms	Estimated Variance Component
Items (i)	9	1.11746	.08413
Judges(j)	6	1.19524	.06667
ij	54	.52857	.52857

r_{kk} = Reliability of average ratings for all judges = 0.53

$\bar{r}_{11}$ = Reliability of ratings for a single judge = 0.14

$\hat{\sigma}^2\,\bar{x}$ = Estimated variance of average ratings = $\hat{\sigma}^2$ (i)/n_i + $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$(ij)/$n_i n_j$ = 0.025

$\hat{\sigma}\,\bar{x}$ = .160

$\hat{\sigma}^2\,\bar{x}_r$ = Estimated variance of ratings over sample of raters = $\hat{\sigma}^2$(j) +$\hat{\sigma}^2$ (ij)/n_i = 0.002

$\hat{\sigma}\,\bar{x}_r$ = .043

$\hat{\sigma}^2\,\bar{x} \mid j$ = Estimated variance of mean ratings generalizing over samples of 10 sets of items for a fixed set of 7 judges = $\hat{\sigma}^2$(i)/n_i + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .016

$\hat{\sigma}\,\bar{x} \mid j$ = .126

$\hat{\sigma}^2\,\bar{x} \mid i$ = Estimated variance of mean ratings generalizes over samples of 7 raters for a fixed set of items = $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .017

$\hat{\sigma}\,\bar{x} \mid i$ = 0.131

Table 5. Summary Statistics for NCB Standard Setting Data

Statistic	Method	
	Angoff	**DSSM**
Average passing score	152.8 (8.1)[1]	174.2 (18.0)
Range of average ratings	144–163.5	143–195
Average time needed to complete ratings	126.5 mins (20.0 mins)	67.3 (11.9 mins)
Average **confidence** ratings:		
Method will accurately separate candidates into certify/ don't certify groups (1 = little confidence, 2 = somewhat confident, 3 = confident, 4 = very confident, 5 = extremely confident)	3.1 (1.1)	3.3 (0.5)
Correlation of ratings [2] with **importance** validity scale ratings (1 = of little importance, 2 = medium importance, 3 = very important, 4 = critically important)	.19	.16
Correlation of ratings with **critical consequences** validity scale ratings (1 = no harm, 2 = minimal harm, 3 = moderate harm, 4 = considerable harm)	.04	.24
Correlation of ratings with item p values	-.01	.18
Correlation of ratings with item r_{pbis}	-.25	-.10

[1]Numbers in parentheses are standard deviation

[2]Correlations based on ratings and/or item statistics summed for 10- item sets

Table 6. NCB Generalizability Analyses for Angoff Method

Source	df	ms	Estimated Variance Component
Items (i)	24	.48	.03414
Judges (j)	5	3.04	.11079
ij	120	.27	.27421

r_{kk} = Reliability of average ratings for all judges = 0.43

$\bar{r}_{11}$ = Reliability of ratings for a single judge = .11

$\hat{\sigma}^2\ \bar{x}$ = Estimated variance of average ratings = $\hat{\sigma}^2$ (i)/n_i + $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$(ij)/$n_i n_j$ = .022

$\hat{\sigma}\ \bar{x}$ = .148

$\hat{\sigma}^2\ \bar{x}_r$ = Estimated variance of ratings over sample of raters = $\hat{\sigma}^2$(j) +$\hat{\sigma}^2$ (ij)/n_i = .122

$\hat{\sigma}\ \bar{x}_r$ = 0.349

$\hat{\sigma}^2\ \bar{x} \mid j$ = Estimated variance of mean ratings generalizing over samples of 25 sets of items for a fixed set of 6 judges = $\hat{\sigma}^2$(i)/n_i + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .003

$\hat{\sigma}\ \bar{x} \mid j$ = 0.057

$\hat{\sigma}^2\ \bar{x} \mid i$ = Estimated variance of mean ratings generalizing over samples of 6 raters for a fixed set of items = $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .020

$\hat{\sigma}\ \bar{x} \mid i$ = 0.142

Table 7. NCB Generalizability Analyses for DSSM

Source	df	ms	Estimated Variance Component
Items (i)	24	1.59	.08317
Judges (j)	5	14.98	.55559
ij	120	1.09	1.09168

r_{kk} = Reliability of average ratings for all judges = 0.31

$\bar{r}_{11}$ = Reliability of ratings for a single judge = 0.06

$\hat{\sigma}^2\ \bar{x}$ = Estimated variance of average ratings = $\hat{\sigma}^2$ (i)/n_i + $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$(ij)/$n_i n_j$ = 0.103

$\hat{\sigma}\ \bar{x}$ = .321

$\hat{\sigma}^2\ \bar{x}_r$ = Estimated variance of ratings over sample of raters = $\hat{\sigma}^2$(j) +$\hat{\sigma}^2$ (ij)/n_i = .599

$\hat{\sigma}\ \bar{x}_r$ = 0.747

$\hat{\sigma}^2\ \bar{x} \mid j$ = Estimated variance of mean ratings generalizing over samples of 25 sets of items for a fixed set of 6 judges = $\hat{\sigma}^2$(i)/n_i + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .011

$\hat{\sigma}\ \bar{x} \mid j$ = 0.103

$\hat{\sigma}^2\ \bar{x} \mid i$ = Estimated variance of mean ratings over samples of 6 raters for a fixed set of items = $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .099

$\hat{\sigma}\ \bar{x} \mid i$ = .316

Table 8. Summary Statistics for ABHP Standard Setting Data

Statistics	Angoff	DSSM
Average passing score	91.9 (7.5)[1]	101.7 (6.5)
Range of average ratings	83.2-108.2	89.0-112.0
Average time needed to complete ratings	90.0 mins (32.8 mins)	67.7 mins (22.1 mins)
Average **confidence** rating:		
Method will accurately separate candidates into certify/don't certify groups (1 = little confidence, 2 = somewhat confident, 3 = confident, 4 = very confident, 5 = extremely confident)	2.8 (0.8)	3.0 (0.5)
Correlation of ratings[2] with item p-values	.38	.20
Correlation of ratings with item r_{pbis}	.33	.28

[1]Numbers in parentheses are standard deviations

[1]Correlations are based on ratings and item statistics summed for 10-item sets

Table 9. ABHP Generalizability Analyses for Angoff Method

Source	df	ms	Estimated Variance Component
Items (i)	14	.144457	-.01643
Judges (j)	8	3.85033	.23719
ij	112	.29248	.29248

$\hat{\sigma}^2\ \bar{x}$ = Estimated variance of average ratings = $\hat{\sigma}^2$ (i)/n_i + $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$(ij)/$n_i n_j$ = 0.27

$\hat{\sigma}\ \bar{x}$ = .166

$\hat{\sigma}^2\ \bar{x}_r$ = Estimated variance of ratings over sample of raters = $\hat{\sigma}^2$(j) +$\hat{\sigma}^2$ (ij)/n_i = 0.257

$\hat{\sigma}\ \bar{x}_r$ = 0.507

$\hat{\sigma}^2\ \bar{x} \mid j$ = Estimated variance of mean ratings generalizing over samples of 15 sets of items for a fixed set of 9 judges = $\hat{\sigma}^2$(i)/n_i + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .001

$\hat{\sigma}\ \bar{x} \mid j$ = 0.033

$\hat{\sigma}^2\ \bar{x} \mid i$ = Estimated variance of mean ratings generalizing over samples of 9 raters for a fixed set of items = $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .029

$\hat{\sigma}\ \bar{x} \mid i$ = 0.169

Table 10. ABHP Generalizability Analyses for DSSM

Source	df	ms	Estimated Variance Component
Items (i)	14	.95238	-.00093
Judges (j)	8	2.80000	.12262
ij	112	.96071	.96071

$\hat{\sigma}^2\ \bar{x}$ = Estimated variance of average ratings = $\hat{\sigma}^2$ (i)/n_i + $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$(ij)/$n_i n_j$ = 0.21

$\hat{\sigma}\ \bar{x}$ = .143

$\hat{\sigma}^2\ \bar{x}_r$ = Estimated variance of ratings over sample of raters = $\hat{\sigma}^2$(j) +$\hat{\sigma}^2$ (ij)/n_i = 0.187

$\hat{\sigma}\ \bar{x}_r$ = 0.432

$\hat{\sigma}^2\ \bar{x} \mid j$ = Estimated variance of mean ratings generalizing over samples of 15 sets of items for a fixed set of 9 judges = $\hat{\sigma}^2$(i)/n_i + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .007

$\hat{\sigma}\ \bar{x} \mid j$ = 0.084

$\hat{\sigma}^2\ \bar{x} \mid i$ = Estimated variance of mean ratings generalizing over samples of 9 raters for a fixed set of items = $\hat{\sigma}^2$ (j)/n_j + $\hat{\sigma}^2$ (ij)/$n_i n_j$ = .021

$\hat{\sigma}\ \bar{x} \mid i$ = 0.144